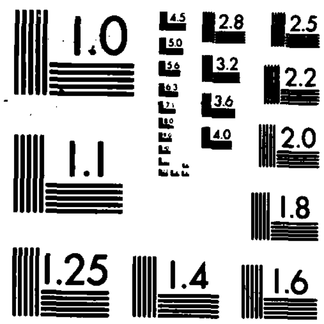
AD-A174 520 SAMPLING PLANS EXCLUDING CONTIGUOUS UNITS(U) ILLINOIS UNIV AT CHICAGO CIRCLE STATISTICAL LAB A S HEDAYAT ET AL. 15 AUG 86 86-13 AFOSR-TR-86-2061 1/1

UNCLASSIFIED F49620-85-C-0008 F/G 12/1 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AFOSR-TR- 86-2061

(2)

THE UNIVERSITY OF ILLINOIS AT CHICAGO

SAMPLING PLANS EXCLUDING CONTIGUOUS UNITS

by

A.S. Hedayat, C.R. Rao and J. Stufken

...ity of Illinois at Chicago and University of Pittsburgh

DEPARTMENT OF MATHEMATICS, STATISTICS, AND COMPUTER SCIENCE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

AD-A174528

REPORT DOCUMENTATION PAGE

- **1a. REPORT SECURITY CLASSIFICATION:** UNCLASSIFIED
- **1b. RESTRICTIVE MARKINGS:**
- **2a. SECURITY CLASSIFICATION AUTHORITY:** N/A
- **2b. DECLASSIFICATION/DOWNGRADING SCHEDULE:** N/A
- **3. DISTRIBUTION/AVAILABILITY OF REPORT:** Approved for public release; distribution unlimited
- **4. PERFORMING ORGANIZATION REPORT NUMBER(S):**
- **5. MONITORING ORGANIZATION REPORT NUMBER(S):** AFOSR-TR- 86-2061
- **6a. NAME OF PERFORMING ORGANIZATION:** University of Illinois at Chicago
- **6b. OFFICE SYMBOL (If applicable):**
- **7a. NAME OF MONITORING ORGANIZATION:** AFOSR/NM
- **6c. ADDRESS (City, State and ZIP Code):** P.O. Box 4348, Chicago, IL 60680
- **7b. ADDRESS (City, State and ZIP Code):** Bldg. 410, Bolling AFB, DC 20332-6448
- **8a. NAME OF FUNDING/SPONSORING ORGANIZATION:** AFOSR
- **8b. OFFICE SYMBOL (If applicable):** NM
- **9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER:** AFOSR 85-0320
- **8c. ADDRESS (City, State and ZIP Code):** Bldg. 410, Bolling AFB, DC 20332-6448
- **10. SOURCE OF FUNDING NOS.**

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.
6.1102F	2304	A5	

- **11. TITLE (Include Security Classification):** Sampling Plans Excluding Contiguous Units
- **12. PERSONAL AUTHOR(S):** A.S. Hedayat, C.R. Rao, J. Stufken
- **13a. TYPE OF REPORT:** Technical
- **13b. TIME COVERED:** FROM ______ TO ______
- **14. DATE OF REPORT (Yr., Mo., Day):** August 15, 1986
- **15. PAGE COUNT:** 16
- **16. SUPPLEMENTARY NOTATION:**
- **17. COSATI CODES**

FIELD	GROUP	SUB. GR.

- **18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number):** Simple Random Sampling; Circular Serial Correlations; Balanced Sampling

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

We consider fixed sized sampling plans for which the second order inclusion probabilities are zero for pairs of contiguous units and constant for pairs of non-contiguous units. A practical motivation for the use of such plans is pointed out and a statistical condition is identified under which these plans are more efficient than the corresponding simple random sampling plans. Results on the existence and construction of these plans are obtained.

- **20. DISTRIBUTION/AVAILABILITY OF ABSTRACT:** UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS ☐
- **21. ABSTRACT SECURITY CLASSIFICATION:** UNCLASSIFIED
- **22a. NAME OF RESPONSIBLE INDIVIDUAL:** Major Brian W. Woodruff
- **22b. TELEPHONE NUMBER (Include Area Code):** (202) 767-5027
- **22c. OFFICE SYMBOL:** NM

DD FORM 1473, 83 APR   EDITION OF 1 JAN 73 IS OBSOLETE.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

SAMPLING PLANS EXCLUDING CONTIGUOUS UNITS

by

A.S. Hedayat, C.R. Rao and J. Stufken

University of Illinois at Chicago and University of Pittsburgh

Statistical Laboratory Technical Report

No. 86-13

August 1986

Research of A.S. Hedayat and J. Stufken is supported by Grant No. AFOSR 85-0320

Research of C.R. Rao is supported by Grant No. AFOSR F49620-85-C-0008

SAMPLING PLANS EXCLUDING CONTIGUOUS UNITS

By

A.S. Hedayat[1], C.R.Rao[2] and J.Stufken[1]

University of Illinois at Chicago and University of Pittsburgh

Abstract

This document

We consider fixed size sampling plans for which the second order inclusion probabilities are zero for pairs of contiguous units and constant for pairs of non-contiguous units. A practical motivation for the use of such plans is pointed out and a statistical condition is identified under which these plans are more efficient than the corresponding simple random sampling plans. Results on the existence and construction of these plans are obtained.

Key Words: Simple random sampling; Circular serial correlations; Balanced sampling

1 Research sponsored by Grant No. AFOSR 85-0320

2 Research sponsored by Grant No. AFOSR F49620-85-C-0008

1. INTRODUCTION AND BACKGROUND

A typical survey sampling set up consists of a population of N labelled units with a value y_i attached to the unit labelled i for i=1,..,N. One of the problems is to estimate $\zeta = (\sum y_i)/N$, the mean value in the population, by observing the y-values on a subset of units in the population. We call

$$d: = \{(s_j, p_j) , j=1,...,b\} \quad , \tag{1.1}$$

where s_j's are subsets of units and $p_j > 0$ is the probability of selection of the subset s_j such that $\sum p_j = 1$, a sampling plan. The corresponding set $\{s_j, j = 1,..., b\}$ is called the support of the sampling plan (SSP). In order to be able to estimate ζ unbiasedly we will assume that $\bigcup s_j = \{1,2,.., N\}$. A sample is a set of labels $s \in$ SSP and the associated y values:

$$X_s = \{(i, y_i), i \in s\} \quad . \tag{1.2}$$

On occasion we will refer to the set s itself as a sample. Let $f(X_s)$ be an unbiased estimate of ζ with variance

$$V_d (f(X_s)) = E (f(X_s))^2 - \zeta^2 = \sum p_j f(X_j)^2 - \zeta^2, \tag{1.3}$$

which depends on the structure of the support and the associated probabilities of the sampling plan d. We may choose $V_d(\cdot)$ as the objective function in preferring one sampling plan to another; the ideal case being when there exists a d with the smallest value for $V_d(\cdot)$.

Throughout this paper our choice for $f(X_s)$ will be the well known Horvitz-Thompson (HT) estimator which has certain desirable statistical properties. Within this framework the only relevant features of the sampling plan are the first and second order inclusion probabilities. A popular choice is to select a sampling plan which has both its first and its second order inclusion probabilities equal. Until 1963 a simple random sampling plan was the only known plan to achieve this. Chakrabarti (1963) initiated an interesting line of work by noticing that the first and second order inclusion probabilities remain fixed after deleting certain samples from the support of a simple random sampling plan, as long as the remaining samples form a BIB design based on N varieties and all samples have an equal probability of selection. Wynn (1977) and Hedayat (1979) observed as a generalization that the support of any BIB design can be used as the support of a sampling plan that achieves the same first and second order inclusion probabilities as a simple random sampling plan. In this case the probabilities of selection should be taken proportional to the frequencies of the blocks in the BIB design. Fienberg and Tanur (1986) also discuss overlaps between design theory and sampling theory.

Although sampling plans with constant second order inclusion probabilities may be appealing, there are situations where their implementation is not advisable. We will now describe such a situation, which can arise naturally in practice, and suggest an alternative plan for it.

In some situations it is conceivable that we are provided with a natural ordering of the N units. As examples one can think of an ordering in time or space. It will in such situations be undersirable to select contiguous units if they provide us with "similar" information. It seems then more reasonable to select a sampling plan that excludes such contiguous units, i.e. a plan for which the second order inclusion probabilities corresponding to pairs of contiguous units ar zero. In the remainder of this section we will give a mathematical model for this set up, while section 2 will discuss the construction of plans that exclude contiguous units. The final section discusses the results and some alternative approaches.

Assume that the N units are arranged in a circular way with i and (i+1) mod N as contiguous units, so that N and 1 are also considered as contiguous. Besides ζ we will need the following parameters of the y values:

$$\sigma^2 = ((y_1 - \zeta)^2 + \cdots + (y_N - \zeta)^2)/N,$$

$$\rho_1 \sigma^2 = ((y_1 - \zeta)(y_2 - \zeta) + \cdots + (y_N - \zeta)(y_1 - \zeta))/N,$$

where ρ_1 is the first order circular serial correlation.

Let us consider a sampling plan for which each sample in the support contains n units and for which the second order inclusion probabilities are zero for pairs of contiguous units and constant for pairs of non-contiguous units. Notice that this also implies that the first order inclusion probabilities are equal for all units. We will refer to such a sampling plan as a balanced sampling plan without contiguous units. since the first and second order inclusion probabilities for such a sampling plan are

$$\pi_i = n/N, \quad i \in \{1,.., N\} \quad ,$$

and

$$\pi_{ij} = n(n-1)/N(N-3)), i \neq j \in 1,...,N, \text{ i and j not contiguous}$$

respectively, it is easily seen that the variance of the HT estimator of ζ is equal to

$$V_1 = \frac{\sigma^2}{n}\left(1 - \frac{(1+2\rho_1)(n-1)}{N-3}\right). \tag{1.4}$$

If we recall that the variance of the HT estimator of ζ in a simple random sampling plan of size n is given by

$$V_2 = \frac{\sigma^2}{n} \; \frac{N-n}{N-1} \quad , \tag{1.5}$$

we see that this latter plan is less efficient than a balanced sampling plan without contiguous units if $\rho_1 > -1/(N-1)$.

It is obvious that the value of ρ_1 depends on the labelling of the units. It is interesting to observe that for any situation, there exists a labelling such that the corresponding ρ_1 value satisfies $\rho_1 \geq -1/(N-1)$. In other words, for any situation, if we use an appropriate labelling a balanced sampling plan without contiguous units will be at least as efficient as a simple random sampling plan. To see this, denote by π all permutations on $\{1,..., N\}$.

Then

$$\frac{1}{N!} \sum_{\pi \in \Pi} \sum_{i=1}^{N} (y_{\pi(i)} - \zeta)(y_{\pi(i+1)} - \zeta)$$

$$= \frac{1}{N-1} \sum_{i_1 \neq} \sum_{i_2} (y_{i_1} - \zeta)(y_{i_2} - \zeta) = -\frac{N\sigma^2}{N-1} ,$$

where $\pi(N+1)$ is by convention the same as $\pi(1)$.

Hence there exists a $\pi_0 \in \Pi$ for which

$$\sum_{i=1}^{N} (y_{\pi_0(i)} - \zeta)(y_{\pi_0(i+1)} - \zeta)/N \geq -\sigma^2/(N-1).$$

The claim follows easily from this if we use a relabelling of the units induced by π_0.

An important question is that of finding an estimate of the expression for V_1 in (1.4). The structure of the support of a balanced sampling plan without contiguous units does not allow us to find an unbiased estimate. One possible option to deal with this problem, as is done in situations similar to this one, is to postulate a model for the y_i's that is consistent with the assumed relation between contiguous units and such that V_1 can be estimated unbiasedly under this super population model. While this is an interesting approach we will instead follow a more traditional approach by approximating those quadratic terms in V_1 which can not be estimated unbiasedly. To do this we will use our assumption that contiguous units provide us with similar information.

The terms in V_1 which prevent us from estimating it unbiasedly are those of the form $y_i y_{i+1}$, $i = 1,..., N$. (It is understood that y_{N+1} denotes y_1, y_{N+2} denotes y_2 etc.) Two possible approximations of $y_i y_{i+1}$ are given by y_i^2 and $\frac{1}{2} y_i(y_i + y_{i+2})$. It is easy to verify that the corresponding approximations of V_1, say V_{11} and V_{12} respectively, can be estimated unbiasedly by

$$\hat{V}_{11} = \frac{N-3n}{n^2 N} \Big(\sum_{i\in s} y_i^2 - \frac{1}{n-1} \sum_{\substack{i_1,i_2 \\ i_1 \neq i_2}} \sum_{\in s} y_{i_1} y_{i_2}\Big)$$

and

$$\hat{V}_{12} = \frac{1}{n^2 N}\Big((N-2n)\sum_{i\in s} y_i^2 - \frac{N-3n}{n-1} \sum_{\substack{i_1,i_2 \\ i_1 \neq i_2}} \sum_{\in s} y_{i_1} y_{i_2} - \frac{n(N-3)}{n-1} \sum_{\substack{i \\ i,i+2 \in s}} y_i\, y_{i+2}\Big)$$

In determining which of these two estimates deserves preference, additional knowledge about the interrelationship between contiguous units will be needed.

2. MATHEMATICAL ASPECTS OF BALANCED SAMPLING PLANS WITHOUT CONTIGUOUS UNITS.

In this section we will study the existence and construction of balanced sampling plans without contiguous units. Throughout we will denote the units by 1, 2,..., N, with i and (i+1) mod N as contiguous units. All samples in the support of the sampling plans will be of fixed size n. The following is an example of a balanced sampling plan without contiguous units for N = 9 and n = 3.

s_j:	{1,3,6}	{1,4,8}	{1,5,7}	{2,4,7}	{2,5,9}	{2,6,8}	{3,5,8}	{3,7,9}	{4,6,8}
p_j:	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9

In this example we see that the first and second order inclusion probabilities are given by

$$\pi_i = 1/3 \quad i \in \{1,..., 9\}$$

and, with $i_1 \neq i_2$

$$\pi_{i_1 i_2} = \begin{cases} 1/9 & \text{if } i_1 \text{ and } i_2 \text{ are non-contiguous} \\ 0 & \text{otherwise} \end{cases}$$

respectively.

Two questions that arise immediately are the following. Given N and n, how many samples without contiguous units do exist, and if all of them have an equal probability of selection do they form a balanced sampling plan without contiguous units? The answers are given in the following theorem.

Theorem 1: For given N and n there are $\frac{N}{N-n}\binom{N-n}{n}$ distinct samples of size n without contiguous units if $N \geq 2n$, and 0 otherwise. If all of them have an equal probability of selection then they form a balanced sampling plan without contiguous units if an only if $n = 2$ and $N \geq 4$.

Proof: If $N < 2n$ it is obvious that there are no samples without contiguous units. If $N \geq 2n$ it is well known that there are $\binom{N-n+1}{n}$ *such samples if we do not consider* 1 and N as contiguous units. (see e.g. Hall (1986)). The number of samples among these that contain both units 1 and N is equal to the number of samples of size n-2 without contiguous units based on 3, ..., N-2, where 3 and N-2 are not considered as contiguous units. By the above result this number equals $\binom{N-n-1}{n-2}$. Hence, considering 1 and N as contiguous units there are $\binom{N-n+1}{n} - \binom{N-n-1}{n-2} = \frac{N}{N-n}\binom{N-n}{n}$ samples without contiguous units.

If all these samples are given an equal probability of selection, assuming $N \geq 2n$, then it is obvious that if $n = 2$ and $N \geq 4$ this leads to a balanced sampling plan without contiguous units. However for $n \geq 3$, $N \geq 2n$ this is no longer true. Let $p: = (\frac{N}{N-n}\binom{N-n}{n})^{-1}$, the probability of selection for each sample. There are $\binom{N-n-2}{n-2}$ samples without contiguous units which contain both the units 1 and 3, while there are $\binom{N-n-3}{n-2}$ such samples which contain both 1 and 4. Therefore $\pi_{13} = n(n-1)/N(N-n-1)$ and $\pi_{14} = n(n-1)(N-2n)/N(N-n-1)(N-n-2)$. Thus $\pi_{13} \neq \pi_{14}$ for $n \geq 3$, which concludes the proof of Theorem 1. □

The result of Theorem 1 tells us that for the construction of a balanced sampling plan without contiguous units we will have to select a subset of all possible samples without contiguous units as its support and/or change the probabilities of selection for the samples. But even with these manipulations it may not be possible to construct such a sampling plan, as the following theorem shows.

Theorem 2: A necessary condition for the existence of a balanced sampling plan without contiguous units based on N units in samples of size n is given by $N \geq 3n$ if $n \geq 3$.

Proof: If possible let $d = \{(s_j, p_j): j = 1,..., b\}$ be a balanced sampling plan without contiguous units with $n \geq 3$ and $N < 3n$. For simplicity define for $i \in \{1,..., N\}$, $\lambda_1 = \sum_{s_j \ni i} p_j$ and for $i_1 \neq i_2 \in \{1,..., N\}$, where i_1 and i_2 are non-contiguous, $\lambda_2 = \sum_{s_j \ni i_1 i_2} p_j$.

First observe that $\lambda_2 (N-3) = \lambda_1 (n-1)$. This can be seen as follows:

$$\lambda_2(N-3) = \sum_{i=3}^{N-1} \sum_{s_j \ni 1,i} p_j = \sum_{s_j \ni 1} p_j \sum_{\substack{i \in s_j \\ i \neq 1}} 1 = (n-1) \sum_{s_j \ni 1} p_j = \lambda_1(n-1). \tag{2.1}$$

Next observe that $\lambda_1 > 2\lambda_2$. Since $\sum_{s_j \ni 3,4} p_j = 0$ we see

$$\lambda_1 = \sum_{s_j \ni 1} p_j \geq \sum_{s_j \ni 1,3} p_j + \sum_{s_j \ni 1,4} p_j = 2\lambda_2. \tag{2.2}$$

If $\lambda_1 = 2\lambda_2$ then by (2.1) we see $N = 2n+1$. The only sample without contiguous units that contains both 1 and 4 is then $\{1,4,6,..., 2n\}$, while the only sample which contains both 2 and 2n is $\{2,4,6,.., 2n\}$. Both must have a probability of selection equal to λ_2. But then, since both contain 4 and 6 we see

$$\lambda_2 = \sum_{s_j \ni 4,6} p_j \geq 2\lambda_2,$$

a contradiction. Hence, $\lambda_1 \neq 2\lambda_2$, which together with (2.2) implies that $\lambda_1 > 2\lambda_2$, as claimed.

Now consider all those samples which contain at least one of the units 1, 2, 3, or 4. These samples will contain exactly two or one of these four units. Without loss of generality let $s_1, \ldots, s_{t_1}$ be the samples that contain two of these units and $s_{t_1+1}, \ldots, s_{t_2}$ those that contain one of them. It is a simple exercise to show that $\sum_{j=1}^{t_1} p_j = 3\lambda_2$ and $\sum_{j=t_1+1}^{t_2} p_j = 4\lambda_1 - 6\lambda_2$.

Since all other units that appear in these t_2 samples are from $\{5,6,\ldots, N\}$ we see that

$$\lambda_1 (N-4) = \sum_{i=5}^{N} \sum_{s_j \ni i} p_j \geq 3\lambda_2 (n-2) + (4\lambda_1 - 6\lambda_2)(n-1) = 4\lambda_1 (n-1) - 3\lambda_2 n. \quad (2.3)$$

Using (2.1) we see that

$$4\lambda_1 (n-1) = 3\lambda_1 (n-1) + \lambda_2 (N-3). \quad (2.4)$$

From (2.4) and (2.3) we obtain

$$\lambda_1 (N-3n-1) \geq \lambda_2 (N-3n-3).$$

Since $N < 3n$ and $\lambda_1 > 2\lambda_2$ we obtain now

$$\lambda_2 (N-3n-3) \leq \lambda_1 (N-3n-1) < 2\lambda_2 (N-3n-1),$$

or $$N > 3n-1,$$

which contradicts $N < 3n$. Hence it must be that $N \geq 3n$, which establishes Theorem 2. □

The next result gives a very general construction for the desired sampling plans. Its generality will also be underlined in Theorem 4.

Theorem 3: Assume that there are t subsets of {1, 2, ..., N} of cardinality n, such that the collection of all tn(n-1) differences modulo N in these t sets contains the elements 2, ..., N-2 equally often, while 1 and N-1 do not appear at all. Then there exists a balanced sampling plan without contiguous units based on 1, 2, ..., N and with samples of size n.

Proof: Develop each of the t sets modulo N, i.e. replace each set, say $(\alpha_1, \ldots, \alpha_n)$ by the N sets $(\alpha_1+1, \ldots, \alpha_n+1), \ldots, (\alpha_1+N, \ldots, \alpha_n+N)$, where the addition is modulo N. The distinct sets obtained in this way form the support of the sampling plan. The probability of selection p_j of a sample s_j in the support is taken proportional to the frequency with which s_j appears in the t N sets above. This gives the desired sampling plan. □

We will call two samples s and s′ cyclicly equivalent if s′ is one of the samples obtained by developing s modulo N. We will say that a sampling plan d has a cyclicly generated support if with $s \in SSP$ it holds that $s' \in SSP$ for any s′ which is cyclicly equivalent with s. We will call d a cyclic balanced sampling plan without contiguous units if it is a balanced sampling plan without contiguous units with a cyclicly generated support and all cyclicly equivalent samples have the same probability of selection. The plans constructed through Theorem 3 are always cyclic balanced sampling plans without contiguous units. The t initial subsets in this construction will be called the generator samples of the sampling plan. As examples of this construction we use Theorem 3 to construct the desired sampling plans for N = 9, n = 3, for N = 15, n = 4 and for N = 10, n =3

1. N = 9, n = 3. Use {1,3,6} as the generator sample. We obtain the following support:

   {{1,3,6}, {2,4,7}, {3,5,8}, {4,6,9}, {5,7,1}, {6,8,2}, {7,9,3}, {8,1,4}, {9,2,5}}

   The probability of selection for each of these samples is 1/9. Notice that this sampling plan is exactly the one in our earlier example. Also notice that N = 3n for this plan.

2. $N = 15$, $n = 4$. Use the following generator sample: {1,3,6,10}. We obtain the following support:

{{1,3,6,10}, {2,4,7,11}, {3,5,8,12}, {4,6,9,13}, {5,7,10,14}, {6,8,11,15}, {7,9,12,1}, {8,10,13,2}, {9,11,14,3}, {10,12,15,4}, {11,13,1,5}, {12,14,2,6}, {13,15,3,7}, {14,1,4,8}, {15,2,5,9}}. The probability of selection for each of these samples is 1/15.

3. $N = 10$, $n = 3$. Use the following 14 generator samples:

{1,3,5}, {1,3,6}, {1,3,6}, {1,3,6}, {1,3,6}, {1,3,6}, {1,3,6}, {1,3,7}, {1,3,7}, {1,3,7}, {1,3,7}, {1,4,7}, {1,4,7}, {1,4,7}

Developing these 14 samples modulo 10 gives us a support of 40 samples. A sample which is cyclicly equivalent to {1,3,5} has a probability of selection of 1/140, if it is cyclicly equivalent to {1,3,6} this is 6/140, to {1,3,7} it is 4/140, to {1,4,7} it is 3/140.

Many other sampling plans have been constructed through the method in Theorem 3, including some families of sampling plans. We give here one such family for $n = 3$. Let $N = 3t$, $t \geq 3$. A balanced sampling plan without contiguous units can be obtained by using the following $\binom{t-1}{2}$ generator samples.

{1,3,6}, {1,3,9},..., {1,3,3(t-1)}, {1,6,9}, {1,6,12},..., {1,6,3(t-1)}, {1,9,12},..., {1,3(t-2), 3(t-1)}.

It is easy to verify that all differences from {2,3,..., 3t-2} appear exactly t-2 times in the collection of 3(t-1)(t-2) differences modulo 3t of the above generator samples.

It is generally possible to use fewer generator samples than those given above. For example if $N = 6t+3$ the following t generator samples will also give the desired sampling plan.

{1,3,6t}, {1,6,6t-3}, {1,9,6t-6},..., {1,3t,3t+3}

The generality of the construction in Theorem 3 is through a simple obsevation emphasized in the following result. Both sampling plans in it are based on N units and fixed sample size n.

Theorem 4: A balanced sampling plan without contiguous units exists if and only if a cyclic balanced sampling plan without contiguous units exists.

Proof: The sufficiency part is obvious. For the necessity part let $d = \{(s_j, p_j), j=1,\dots, b\}$ be a balanced sampling plan without contiguous units. Define a new sampling plan $d^* = \{(s_j^*, p_j^*), j = 1,\dots, b^*\}$ as follows. For its support develop each of the samples s_j modulo N. If $s_{j_o}^*$ is one of the samples obtained, define $p_{j_o}^*$ as

$$p_{j_o}^* = \frac{1}{N} \sum_{s_j \sim s_{j_o}^*} \alpha_j \, p_j \ ,$$

where the summation is over all those j's for which s_j is cyclicly equivalent to $s_{j_o}^*$ and where α_j denotes the frequency of appearance of $s_{j_o}^*$ after developing s_j. This gives a cyclic balanced sampling plan without contiguous units. □

Notice that the α_j's are often equal to one, but not always. For example developing {1,4,7} modulo 9 gives $\alpha_j = 3$. Further observe that if the probabilities p_j are rational, then d^* can be constructed through Theorem 3. so as a corollary we obtain:

Corollary 1: A balanced sampling plan without contiguous units and with rational probabilities of selection exists if and only if a cyclic balanced sampling plan without contiguous units can be constructed through the method in Theorem 3.

Regarding the existence question of the desired sampling plans the method in Theorem 3 is thus very powerful. This may not be true if our main objective is to find such a sampling plan with the smallest support size.

Given a balanced sampling plan without contiguous units, based on N units in samples of size n, there is an easy way to obtain such a sampling plan based on N units in samples of size $n' \leq n$. If $d = \{(s_j, p_j), j=1,\dots, b\}$ is the plan with samples of size n define $d^* = \{(s_j^*, p_j^*), j=1,\ldots, b^*\}$ as follows. Its support consists of those samples s^* of size n' for which there is an s_j such that $s^* \subset s_j$. The probability of selection p^* of s^* is defined as

$$p^* = \left(\sum_{s_j \supset s^*} p_j \right) / \binom{n}{n'}$$

Then d^* is a balanced sampling plan without contiguous units, based on N units in samples of size n'. so we showed:

Theorem 5: The existence of a balanced sampling plan without contiguous units, based on N units in samples of size n implies the existence of a balanced sampling plan without contiguous units, based on N units in samples of size $n' \le n$.

We point out that there is a simple generalization of this observation for the case that the initial sampling plan contains samples of different sizes.

We conclude this section with the remark that all our results can immediately be translated to the language of incomplete block designs. several other results, e.g. a Fisher-type inequality, a lower bound for the support size through a Mann-type inequality, can be obtained in this context.

3. SUMMARY AND DISCUSSION

In the previous two sections we have given a motivation for the possible use of balanced sampling plans without contiguous units and have studied some of their mathematical aspects. We have demonstrated that there are situations in which it is undesirable to collect information from contiguous units, situations for which balanced sampling plans without contiguous units are more efficient than simple random sampling plans for estimating the population mean by the HT estimator. We have also addressed the problem of estimating the variance of the HT estimator under balanced sampling plans without contiguous units. In section 2 we established some useful results or the existence and construction of the desired sampling plans.

An obvious generalization to our problem is to consier situations for which it is not only desirable to exclude (first order) contiguous units from the sampling plan, but also higher order contiguous units. parameters that become of interest for such cases are the higher order circular serial correlations, e.g. ρ_2 the second order circular serial correlation defined by

$$\rho_2 \, \sigma^2 = ((y_1 - \zeta)\,(y_3 - \zeta) + \ldots + (y_N - \zeta)\,(y_2 - \zeta))/N.$$

We would like to point out that if $N = (m+1)n$ and contiguous units of order less than or equal to m have to be excluded, the situation corresponds to that of systematic sampling.

Finally we like to point out the possiblity of studying the problem addressed in this paper by looking at other estimators than the HT estimator for the population mean. This could be done instead of manipulating the sampling plan or in addition to such manipulations. For more details on this we refer the reader to Rao (1975).

REFERENCES

Chackrabarti, M.C., (1963). On the use of incidence matrices in sampling from finite populations. J. Indian Statist. Assoc. 1, 78-85.

Fienberg, S.E. and Tanur, J.M. (1986). Experimental and sampling structures: parallels diverging and meeting. Tech. Rep. No. 374, Dept. of Statistics, Carnegie-Mellon University.

Hedayat, A. (1979). Sampling designs with reduced support sizes. In: Optimizing methods in statistics (J. Rustagi, ed.), 273-288, Academic press.

Rao, C.R. (1975). Some problems of sample surveys. Adv. Appl. Prob (supplement). 7, 50-61.

Wynn, H.P. (1977). Convex sets of finite population plans. Ann. Statist. 5, 414-418.

END

12-86

DTIC